La plus grande valeur propre de matrices de covariance empirique

Sandrine PÉCHÉ*

Dans cet article nous expliquons brièvement l'intérêt de l'étude des valeurs propres extrêmes de matrices aléatoires hermitiennes de grande taille. Nous donnons ensuite les grandes lignes des méthodes d'étude de ces propriétés fines du spectre.

Matrices de covariance empirique

Une motivation statistique

Bob est statisticien et travaille dans le département de recherche d'une grande banque. Il veut modéliser l'évolution du marché pour les 10 prochaines années, afin de pouvoir faire des prévisions sur les cours d'un grand nombre d'actions. Pour ce faire, il choisit de représenter dans un tableau les rendements d'un grand nombre N d'actions sur un grand nombre p de jours. Typiquement p et N peuvent être d'ordre 10^4. On peut montrer qu'en utilisant un modèle aléatoire, on obtient une bonne représentation de ces rendements. En effet, les variations des cours sont particulièrement désordonnées, les marchés étant très sensibles à de multiples facteurs, qui peuvent varier suivant les titres et au cours du temps. Ainsi Bob choisit de modéliser l'évolution des rendements dans une matrice de taille $N \times p$, $M = [Z_1, \ldots, Z_N]$, où les Z_i sont des vecteurs de taille p représentant les p valeurs des rendements de chaque action. Il fait les hypothèses suivantes :

– les entrées $Z_{ij}, j = 1, \ldots, p$ des vecteurs Z_i sont des variables aléatoires

– elles sont mutuellement indépendantes.

La matrice $M = [Z_1, \ldots, Z_N]$ est ce que l'on appelle une *matrice aléatoire réelle*, dont l'étude a réellement commencé avec [10] en physique nucléaire.

Le problème posé : Afin de donner à ses supérieurs un compte-rendu fiable des données prévisionnelles, sans entrer dans une lecture longuette de tous les chiffres, Bob doit donc trouver un moyen de résumer au mieux l'information dont il dispose, tout en pouvant préciser quelle erreur il commet en la résumant.

Le problème de Bob est en fait un vieux de problème de statistique. Dès les années 1930, Wishart et James ([2]) sont les premiers à considérer le moyen optimal de résumer l'information statistique recueillie en effectuant p mesures sur une population de taille N. Le but est de minimiser les coûts des calculs numériques sur des très grandes matrices. Ce moyen sera par la suite défini par [1] en 1933, donnant naissance à l'analyse par « composantes prin-

* *Institut Fourier, Université Joseph Fourier,*
UMR 5582, BP 74, 38402 St Martin d'Heres Cedex, France.
sandrine.peche@ujf-grenoble.fr

cipales ». Nous allons rappeler et expliquer les grands principes de cette méthode et l'intérêt de l'étude des plus grandes valeurs propres de certaines matrices de « covariance empirique » pour son application.

Matrices de covariance empirique

Commençons par quelques rappels. Soient p et N des entiers fixés avec $p \geqslant N$. Soit M une matrice complexe de dimension $N \times p$. La matrice M^*, de dimension $p \times N$, est définie par $M^*_{ji} = \overline{M_{ij}}$ pour $i = 1,\dots,N$ et $j = 1,\dots,p$.

Définition 1. *Une matrice$X$ est Hermitienne si $X = X^*$. Une matrice Hermitienne X est positive si $\forall V \in \mathbb{C}^N$, $V^*XV \geqslant 0$.*

Typiquement, la matrice $X = MM^*$ est une matrice Hermitienne positive.

D'après le théorème spectral, on sait que la matrice X est diagonalisable dans une base orthonormée B de $\mathbb{C}^N$, $B = (V_1,\dots,V_N)$, et admet N valeurs propres positives $\lambda_1 \geqslant \cdots \geqslant \lambda_N \geqslant 0$. Ceci s'écrit mathématiquement $X = VDV^*$, si V est la matrice $(V_1 \ \cdots \ V_N)$ et $D = \mathrm{diag}(\lambda_1,\dots,\lambda_N)$, avec $XV_i = \lambda_i V_i$, pour $i = 1,\dots,N$, et $VV^* = Id$.

Remarque 2. *On sait aussi que la matrice $W = M^*M$ de taille $p \times p$ admet les mêmes valeurs propres non nulles que la matrice X. Ainsi il existe une matrice U de taille $p \times p$ telle que $UU^* = Id$ et $W = U\mathrm{diag}(\lambda_1,\dots,\lambda_N,0,\dots,0)U^*$.*

Nous donnons maintenant la définition dans un cadre général d'une matrice de covariance empirique complexe.

Soit μ,μ' deux mesures de probabilité sur $\mathbb{R}$.

Définition 3. *Une matrice de covariance empirique complexe est une matrice X_N avec $X_N = \frac{1}{N}MM^*$ si $M = Z + iY$, avec Z,Y de taille $N \times p$ dont les entrées $Z_{i,j}$, (resp.Y_{ij}), $i = 1,\dots,N$, $j = 1,\dots,p$, sont des variables aléatoires indépendantes identiquement distribuées (i.i.d.) de loi μ (resp. μ').*

On supposera toujours que les lois μ et μ' sont centrées et de variance finie indépendante de N. On note alors $\sigma^2 := Var(Z_{11}) + Var(Y_{11})$.

Exemple important : Si les entrées Z_{ij} et Y_{ij} sont des Gaussiennes i.i.d. $\mathcal{N}(0,\sigma^2/2)$, la matrice $X_N = X_N^{LUE}$ est alors dite de l'ensemble de Laguerre unitaire, noté LUE. Le LUE est la loi de cette matrice.

Dans la suite, on note W_N la matrice $\frac{1}{N}M^*M$, naturellement associée à X_N. On note aussi $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_N \geqslant 0$ les valeurs propres ordonnées de X_N et V_i (resp. U_i) les vecteurs propres orthonormés de X_N (resp. W_N) associés.

Le problème de Bob : Supposons que Bob a modélisé les cours par une matrice aléatoire réelle M comme dans la définition 3. Cette modélisation est ici trop simpliste, mais elle va nous permettre d'expliquer les principes de l'étude des matrices de covariance. Bob cherche une approximation de la matrice $M_N = N^{-1/2}M$ par une matrice Y de rang plus petit. Il veut aussi maximiser la qualité de l'approximation, mesurée par la quantité

$$Q(Y) = \frac{\sum_{i=1}^N \sum_{j=1}^p |Y_{ij}|^2}{\sum_{i=1}^N \sum_{j=1}^p |(M_N)_{ij}|^2}.$$

L'analyse par composantes principales résout ce problème. La meilleure approximation de M_N par une matrice de rang 1 (par exemple) est alors la matrice $M_1 = \sqrt{\lambda_1} V_1 U_1^*$ et sa qualité est $Q(M_1) = \dfrac{\lambda_1}{Tr(X_N)}$. La meilleure approximation de rang $k > 1$ est aussi connue et sa qualité s'exprime en fonction des k plus grandes valeurs propres. Revenons à l'approximation de rang 1. Afin de contrôler l'erreur commise, Bob doit déterminer les propriétés de la plus grande valeur propre λ_1 et de la trace de la matrice X_N. Calculer toutes les valeurs propres de X_N pour déterminer ensuite la plus grande, et ce pour chaque réalisation aléatoire est numériquement trop coûteux. Nous allons ici donner d'autres méthodes permettant d'étudier λ_1, et $Tr(X_N)$, quand N est grand.

Comportement global des valeurs propres

Nous allons maintenant obtenir une première borne inférieure pour la plus grande valeur propre (et identifier au passage le comportement limite de $Tr(X_N)$). Pour simplifier, on suppose à partir de maintenant que $p - N$ est un entier fixé. Le nombre de données p est donc comparable à la taille de la population N (*cf.* [3] pour une étude plus générale).

La méthode est la suivante. Etant donné un intervalle borné I, nous allons dénombrer le nombre de valeurs propres qui tombent dans cet intervalle.

Théorème 4. *Soit I un intervalle quelconque. La proportion des valeurs propres de X_N dans I, notée $N_N(I)$, est donnée asymptotiquement par*

$$\lim_{N\to\infty} N_N(I) = \int_I \frac{1}{2\pi\sigma^2\sqrt{x}} \sqrt{4\sigma^2 - x}\, 1_{[0,4\sigma^2]}(x)dx. \qquad (1)$$

Remarque 5. *La fonction intégrée dans (1) est une densité de probabilité. Elle définit la loi dite de Marchenko-Pastur.*

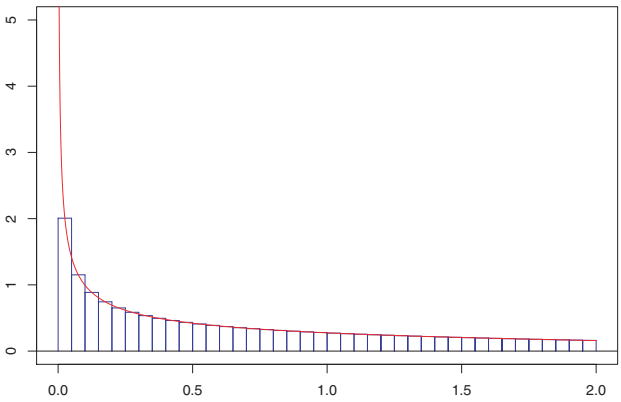


Figure 1 – *Histogramme des valeurs propres et loi de Marchenko-Pastur.*

La Figure 1 montre l'histogramme des valeurs propres d'une matrice de taille 40, et montre l'adéquation avec la densité de la loi de Marchenko et Pastur. Grossièrement, on est « sûr » de trouver, pour N assez grand, des valeurs

propres de X_N dans tout sous intervalle de $I = [0,4\sigma^2]$. On a donc, pour N assez grand, et en dehors d'un ensemble de probabilité nulle, $\limsup_{N\to\infty} \lambda_1 \geqslant 4\sigma^2$.

Remarque 6. *Une preuve consiste à étudier les « moments » d'ordre $k \in \mathbb{N}$ fixé, $\frac{1}{N}\mathbb{E}\left(Tr\left(X_N^k\right)\right) = \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}\lambda_i^k\right)$, et de montrer que ces moments convergent, quand $N \to \infty$ vers le moment d'ordre k de la loi de Marchenko-Pastur, si la variance σ^2 est fixée. En particulier, $\frac{1}{N}Tr(X_N)$ converge vers σ^2.*

Des questions se posent alors naturellement. Par exemple,

– a-t-on $\lim_{N\to\infty} \lambda_1 = 4\sigma^2$?

– Si oui, quel est la vitesse de convergence de λ_1 vers $4\sigma^2$?

– La loi limite de $\lambda_1 - 4\sigma^2$ est-elle centrée ou non ?

Le théorème de Marchenko et Pastur ne nous permet pas de répondre à ces questions. Il concerne en effet un comportement de type global, à savoir les propriétés de toutes les valeurs propres, considérées simultanément. Pour étudier le comportement plus précis de λ_1, nous devons définir des outils qui permettent de différencier mieux les valeurs propres, sans toutefois revenir au calcul de chacune d'entre elles.

L'ensemble de Laguerre unitaire

Pour étudier le comportement de λ_1, nous avons besoin de faire des hypothèses supplémentaires sur la loi des entrées de X_N. La seule connaissance de la variance σ^2 ne semble pas suffire. Nous allons donc nous intéresser plus particulièrement au LUE. Le LUE présente en effet deux particularités, qui ne sont pas vraies en général pour des entrées non Gaussiennes, et qui font que c'est l'ensemble mathématiquement parlant le plus simple.

D'abord, on sait calculer explicitement la densité de probabilité jointe $g : \mathbb{R}_+^N \to \mathbb{R}_+$ des valeurs propres du LUE. Le calcul remonte à [2]. Cette densité est importante car elle permet, *a priori*, de déterminer la fonction de répartition de la plus grande valeur propre. En effet,

$$\begin{aligned} \mathbb{P}(\lambda_1 \leqslant s) &= \mathbb{P}(\text{toutes les valeurs propres sont dans } (-\infty,s]) \\ &= \int_{-\infty}^{s} \cdots \int_{-\infty}^{s} g(x_1,\cdots,x_N) \prod_{i=1}^{N} dx_i. \end{aligned} \tag{2}$$

Reste le calcul de cette intégrale N-dimensionnelle, ce qui n'est pas connu en général. C'est là la deuxième particularité du LUE. On peut explicitement calculer la fonction de répartition, ce qui a été obtenu par Bronk, [4]. Ce calcul utilise une formidable astuce initialement due à Mehta, Gaudin, ([8], chap. 5), qui exprime (2) comme une certaine fonction des polynômes orthogonaux de Laguerre (*cf.* [7]) ! Ceci explique d'ailleurs la dénomination LUE. Ces polynômes étant parfaitement bien connus, on a pu en déduire le comportement asymptotique de λ_1 pour le LUE, que nous donnons maintenant.

On appelle loi de Tracy-Widom, de fonction de répartition F_2, la mesure de probabilité dont la densité est donné par la Figure 2. La définition mathématique de cette loi, plutôt compliquée, est donnée dans [9].

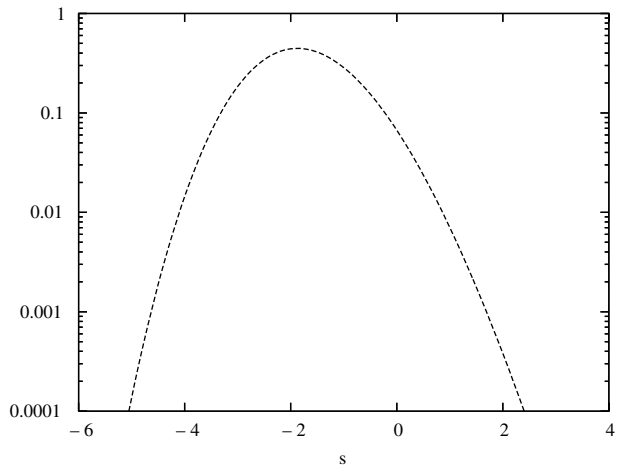


Figure 2 – *Densité de la loi de Tracy-Widom.*

Théorème 7. *[5] Soit $x_o \in \mathbb{R}_-$ fixé et $x \in [x_o, \infty)$. Alors,*

$$\lim_{N\to\infty} \mathbb{P}\left(N^{2/3} \frac{\lambda_1^{LUE} - 4\sigma^2}{4\sigma^2} \leqslant x \right) = F_2(x).$$

Le Théorème 7 répond ainsi aux questions de la section précédente, dans le cas du LUE. D'abord, quand $N \to \infty$, λ_1^{LUE} converge vers $4\sigma^2$. Elle n'a donc pas tendance à se séparer des autres valeurs propres. De plus, elle fluctue autour de $4\sigma^2$ dans des intervalles de longueur typique d'ordre $N^{-2/3}$. Les fluctuations de cette valeur propre autour de $4\sigma^2$ et dans l'échelle typique sont aléatoires et de loi asymptotiquement donnée par la loi de Tracy-Widom, qui n'est pas centrée.

Modèles de matrices plus généraux

Une fois le comportement de λ_1 identifié pour le LUE, on montre que ce comportement est en fait valable pour des matrices X_N plus générales. L'idée de base est la suivante. Soit $A > 0$ fixé (grand), et k le nombre des valeurs propres $\lambda_i > 4\sigma^2 - AN^{-2/3}$. Ecrivons alors $\lambda_i = 4\sigma^2(1 + \chi_i N^{-2/3})$, où χ_i est une variable aléatoire de loi F_i. On obtient

$$\mathbb{E}\left(Tr\left(\frac{X_N}{4\sigma^2} \right)^{tN^{2/3}} \right) = \mathbb{E}\left(\sum_{i=1}^{N} \left(\frac{\lambda_i}{4\sigma^2} \right)^{tN^{2/3}} \right) \simeq \mathbb{E}\left(\sum_{i=1}^{k} \int_{\mathbb{R}} e^{tx} dF_i(x) \right).$$

Si on montre que, pour tout $j \in \mathbb{N}$ et pour tout réels positifs $t_1, \ldots, t_j$, la limite $\lim_{N\to\infty} \mathbb{E}\left(\prod_{i=1}^{j} Tr\left(\frac{X_N}{4\sigma^2} \right)^{t_i N^{2/3}} \right)$ existe et ne dépend que de σ^2, comme dans le cas du LUE, alors on montre « grossièrement » que les plus grandes valeurs propres de X_N ont le même comportement asymptotique que celles du LUE.

Théorème 8. *[6] Si pour tout entier k, $\mathbb{E}\left((M_{11})^{2k+1} \right) = 0$, et $\mathbb{E}\left(|M_{11}|^{2k} \right) \leqslant (Ck)^k$, alors*

$$\lim_{N\to\infty} \mathbb{P}\left(N^{2/3} \frac{\lambda_1^{LUE} - 4\sigma^2}{4\sigma^2} \leqslant x \right) = F_2(x).$$

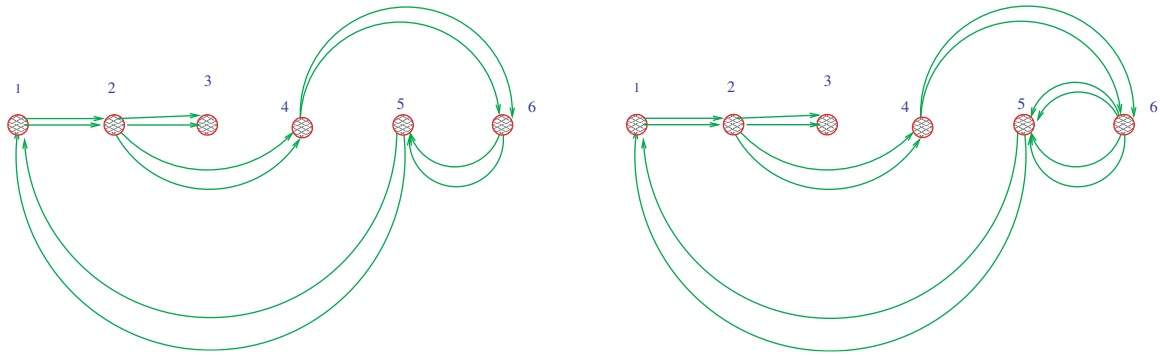
Remarque 9. *Une matrice du LUE satisfait les conditions du Théorème 8.*

Pour esquisser la preuve, supposons que $j = 1$ et que $tN^{2/3} = l_N$ est un entier pair. On développe la trace

$$\mathbb{E}\left(Tr\left(\frac{X_N}{4\sigma^2}\right)^{l_N}\right) = \frac{1}{(4\sigma^2)^{l_N}} \sum_{i_o,\ldots,i_{l_N-1}=1}^{N} \mathbb{E}\left(X_{i_o i_1} X_{i_1 i_2} \cdots X_{i_{l_N-1} i_0}\right)$$

$$= \frac{1}{(4\sigma^2 N)^{l_N}} \sum_{k_1,\ldots,k_{l_N-1}=1}^{p} \sum_{i_o,\ldots,i_{l_N-1}=1}^{N} \mathbb{E}\left(M_{i_o k_1} \overline{M_{i_1 k_1}} \ldots M_{i_{l_N-1} k_{l_N-1}} \overline{M_{i_o k_{l_N-1}}}\right). \tag{3}$$

A chacun des termes de (3), on associe un graphe orienté : on trace les arêtes du sommet i_o vers k_1, de i_1 vers k_1, ..., de i_{l_N-1} vers k_{l_N-1} et de i_o vers k_{l_N-1}. On obtient donc $4s_N$ arêtes reliant des sommets choisis dans $\{1,\ldots,N\}$ ou $\{1,\ldots,p\}$. On regroupe les termes associés à chaque arête orientée et on calcule l'espérance associée. Or, dès qu'une arête apparaît un nombre impair de fois, cette espérance est nulle. On tient donc compte des seuls graphes où chaque arête est parcourue un nombre pair de fois.



(a) Chemin valide (b) Chemin négligeable

Figure 3

L'étape suivante est de montrer que les seuls graphes qui vont avoir une contribution significative sont ceux pour laquelle chaque arête est passée au plus deux fois. Par exemple, le chemin de la Figure 3, où l'arête (6,5) est passée 4 fois, ne sera pas parmi les chemins à comptabiliser. En effet, on choisit beaucoup moins de sommets (puisqu'on en répète) que dans le cas où les arêtes sont « doubles ». La présence du facteur $1/N^{l_N}$ fait alors que ces chemins avec arêtes plus que doubles sont négligeables. Or pour chaque arête passée deux fois, on a $\mathbb{E}\left(M_{ij} M_{ji}\right) = \sigma^2$ ou, les entrées étant complexes non réelles, $\mathbb{E}\left(M_{ij} M_{ij}\right) = 0$. Le résultat final s'exprime donc en fonction de σ^2 seulement, comme dans le cas du LUE. On en déduit ensuite le Théorème 8.

Conclusion

Nous avons donné ici les idées d'une méthode d'étude des plus grandes valeurs propres dans le cas général de matrices aléatoires complexes. Des petites modifications sont à apporter dans le cas de matrices réelles, où les formules sont en fait plus compliquées ! Concluons avec le problème de Bob. En choisissant des modèles de matrices aléatoires un peu plus compliqués (et plus proches de la réalité) que celui présenté ici, par exemple avec des entrées M_{ij} de variances différentes, on peut montrer que les marchés sont en fait très bien représentés par des matrices de très faible rang (au plus 10). C'est un point très interessant pour la gestion d'un portefeuille...

Pour en savoir plus

[1] Hotelling (H.), Analysis of a complex of statistical variables into principal components, *Jour. Educ. Psych.*, 24 : 417–441, (1933).

[2] James (A.), Distributions of matrix variates and latent roots derived from normal samples, *Annals of Mathematical Statistics*, 35 : 475–501, (1964).

[3] Marcenko (V.A.) and Pastur (L.A.), Distribution of eigenvalues for some sets of random matrices, *Math. USSR-Sbornik*, 1 : 457–486, (1967).

[4] Bronk (B.V.), Exponential ensembles for random matrices, *J. Math. Phys,* 6 : 228–237, (1965).

[5] Forrester (P.J.), The spectrum edge of random matrix ensembles, *Nuclear Physics B*, 402 : 709–728, (1993).

[6] Soshnikov (A.), A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices, (2001), Preprint, arXiv : math. PR/0104113 v2.

[7] Szego (G.), *Orthogonal polynomials*, American Mathematical Society, Providence, RI, (1967).

[8] Mehta (M.), *Random matrices,* Academic press, San Diego, second edition, (1991).

[9] Tracy (C.) and Widom (H), Level spacing distributions and the Airy kernel, *Comm. Math. Phys*, **159** : 33–72, (1994).

[10] Wigner (E.) Characteristic vectors bordered matrices with infinite dimensions, *Ann. Math.* **62** : 548–564, (1955).